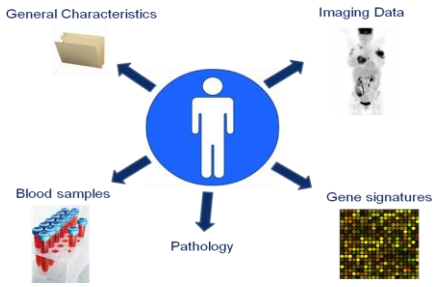
Consensus Position on the Assessment of Quantitative Imaging Biomarkers with Implication to Process Roadmap

The purpose of this paper is to converge the thoughts of otherwise diverse stakeholders who are interested and active in the definition, validation, and qualification of imaging as a biomarker. The approach is to create a series of connected thoughts that present a series of propositions understood and agreed to by the stakeholders so as to reach agreement on the collaborative activities needed to advance specific biomarkers as well as the field in general.

Clinical and Regulatory Relevance

Using patient individual information we want to predict prognosis and the optimal treatment for a specific patient. Towards an individualized treatment instead of the now standard "one size fits all" treatment modalities.



Beyond just diagnostics, images are non-invasive 'biomarkers' capable of providing objective measures of tissue characteristics with numeric readouts capable of being incorporated into phenotyping (for diagnosis and stratification) and/or longitudinal measurements of disease progression (including therapy response assessment). Some stakeholders are primarily interested in phenotyping, some are interested in progression, and some in both.

Phenotyping

Phenotyping is essentially a classification problem, where biomarkers (whether imaging or otherwise) provide features capable of discriminating across classes (also referred to as types or subtypes). The accepted performance assessment method for classifiers are receiver operating characteristic (ROC) curves representing the performance by plotting sensitivity vs. specificity (actually 1-specificity) (often labeled as "true positives" and "false positives"), where the goal is to get area under the curve (AUC) as close as possible to 1 (which indicates a perfect classifier):

Area under the Curve (AUC):

AUC = 1.0: Perfect

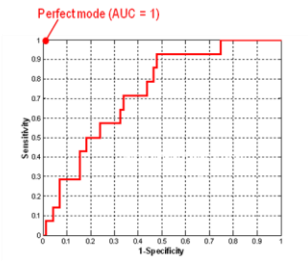
AUC = 0.9: Very Good

AUC = 0.8: Good

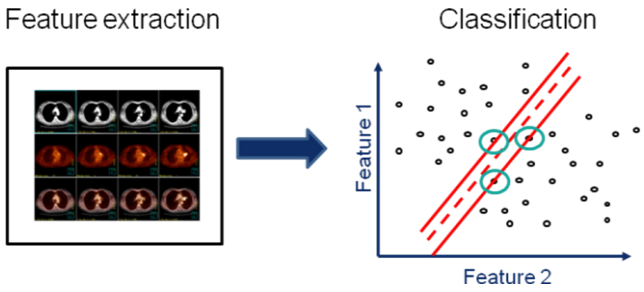
AUC = 0.7: Reasonable

AUC = 0.6: Bad

AUC = 0.5: Random (very bad)

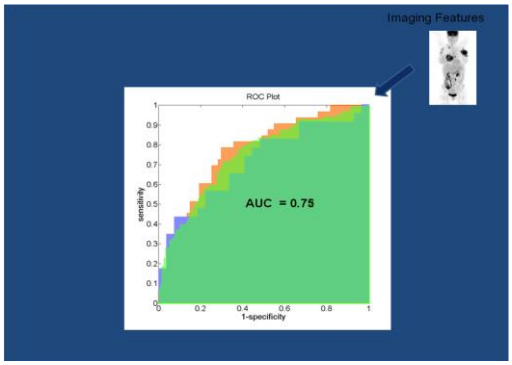


Imaging Biomarkers (as with any other biomarker types) are understood to be the extraction and analysis of at least one and often multiple features mathematically derived from the image data (e.g., volume, density, …). These are viewed as n features which together comprise the imaging phenotype (n is usually small so as not to overfit the available data, but not necessarily only 1).

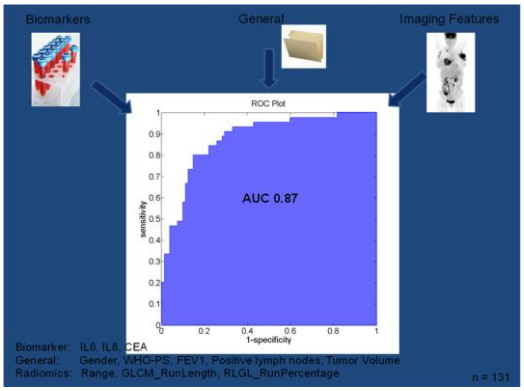


(note that SVM is only one way of approaching classification)

Features with a high effect size large enough to have a high positive predictive value when used in a classifier that is evaluated on a large enough sample population to be statistically similar to the target population are said to be attractive:



Even better is to find features across all available patient data that increase the AUC by exploiting heterogeneous data types with complementary information content with respect to the classification. For example, genotypic and phenotypic measures are generally complementary and add to predictive power:



There can be errors made with any assay. The errors may just add noise in epidemiologic studies or as a covariate in clinical trials, but the tolerance for mis-measurement when specific individual patient management is at stake should be much lower.

The merit of imaging assays that extract one or more quantitative features is demonstrated by positive contributions to AUC after subtracting test-retest reproducibility. The goal is to find features that have an effect size that is not only larger than the variance in the estimates but which is large enough to make an incremental contribution to the AUC as practiced without benefit of the assay.

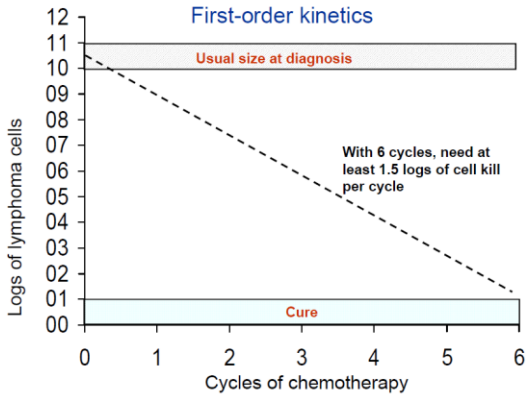
Progression (Therapy response assessment)

Measurement of progression may be more or less difficult than phenotyping at a given time point but is nonetheless a separate problem with its own evaluation issues. In certain cases it may be easier to phenotype at a given time point, if the relative change between time points is small with respect to the variability of the assay. In other cases it may be easier to measure progression, if systematic biases at individual time points have no practical mitigation strategy.

The definition of progression depends on the nature of the pathogenesis. Using cancer tumors as example, the conventional response assessment technique of RECIST has recognized limitations for many subtypes and therapies, including low analytical power per subject in the conventional clinical trial setting as well as inability to measure nascent response rapidly enough for adaptive trials or generally in neoadjuvant settings. In the former case, the result is excessive cost and time in trials; in the latter case, various novel therapeutic concepts cannot be properly evaluated at all.

There are two ways to do better: increase the sensitivity of the current (size) measure, and/or replace the measure with one more correlated to the relevant biology. In any event, the ability of the assay to measure significant change on the same time course of the evolving biology and at a sensitivity well matched to the underlying mechanism which is measured together constitute the performance of the change measure with respect to intended clinical use.

For example, in cytotoxic therapy, early indicators must be tuned to measuring the first “couple” logs of cell kill (that may not yet manifest as a change in tumor size), yet sustained response requires indicators tuned to accurate indicator of size (since the sensitivity at higher log reductions may exceed the capability of even the most sensitive available imaging technology currently):



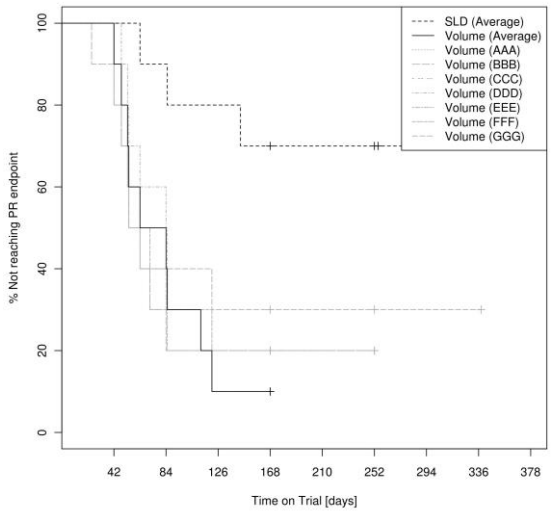
Given that size change only starts to occur after sufficient degree of cell kill, suggesting that modalities that can measure nascent pathogenic processes afford the earliest indicators. But they tend to be focused on one mechanism of action rather than the aggregate of all so they may not be as conclusive as size in the mid- to long-term. With respect to a better size measure, volumetric as opposed to diameter analysis for example may better utilize the available image data which suggests it may be better able to assess the response trend (both in time and slope) in the mid-term. In any event, different features offer promise to reduce either the time or enrollment needed to achieve a given level of statistical power in response studies.

Even though some point out that variance in an imaging method may cancel out across clinical trial arms, it will drive sample sizes and may result in inappropriate termination of therapy in an arm where it is working (or vice versa). Random errors of progression or regression weaken studies. As an illustration, a test wrong 25% of the time contaminates the data markedly.

If there are many subjects from a small number of centers, systematic bias may be avoided by randomization by center. As the number of centers increases, the potential bias is manifest as extra variability. There are many sources of variability in multicenter trials. Arguably, there is even more variability in clinical practice, which is what the trials are trying to guide.

The matter of differential bias or having different bias in various trial arms is even more to be avoided. An example would be patient selection for high and low risk arms based on SUV. As lower SUV also likely corresponds with 'smaller' tumors, partial volume effects may also result in different biases between both groups. There is often interplay between accuracy (or bias per subject group) versus precision.

Ultimately the merit of the progression biomarker is demonstrated by Kaplan-Meier estimator after subtracting test-retest reproducibility:



The goal is to find biomarkers that can serve as surrogate endpoints due to their effect size being not only larger than the variance in the estimates but which is large enough to make an incremental contribution to the AUC as practiced without benefit of the assay.

Implications to Process Roadmap

Inaccuracy and imprecision drive costs and may inappropriately segregate patients to the wrong treatment. As such, it is important to identify both the intended purpose of the assay and to characterize its performance in separate efforts where the performance differs for different purposes. Likewise, the precision of the variability estimate itself depends on the purpose; 1% here or there will not greatly influence a late stage trial. But most clinical trials are not Phase III trials. The real go/no go decisions for therapies occur much earlier, and if imaging will play a role it has to work in a single-arm setting, or be pretty consistent in any small sample setting. For example, a pick-the-winner randomized phase II. Likewise, for individualized patient management the sensitivity due to error may be higher for ethical reasons.

QIBA focuses on characterizing assay performance with respect to specified intended uses. Likewise, it seeks collaborative efforts to reducing the variance and improving accuracy. Precision and reduction of differences in bias between sites is most important, even when bias still remains. A balance is struck on a per team basis as to how much effort is needed to complete characterization of assays vs. the amount of effort to improve them. Likewise, a balance between finding those biomarkers with an effect size that is inherently less sensitive to noise vs. working to optimize biomarkers that may currently have equivocal effect size such that decreasing variance is needed to make it work.

There are two mutually-related but semi-independent lines of activity to pursue this.

One is dominated by physicists and technical personnel, including vendors, to pursue the creation, refinement, and validation of the assay. For each biomarker, the following steps are undertaken:

1. Discovery of potential imaging biomarker (do tools exist or do new assays or other supporting tools need to be developed?)
2. Test / refine imaging performance, PK/PD, toxicology, etc. in preclinical setting.
3. Pursue IDE if necessary, submit IND if necessary, and/or optimize existing platform for new use.
4. Acquire or develop phantom and other controlled condition support material for controlled experimentation and ongoing QC.
5. Define and iteratively refine acquisition, analysis, interpretation, QC, etc. for specific clinical utility, defining and closing "gaps".
6. Assessment of intrinsic scanner variability, minimum detectable change, and other aspects of assay performance in controlled conditions.
7. Assessment of intra- and inter-reader (human-drive and/or computer-alone as indicated by intended use) bias and variance across scanners and centers.

The other activity that presupposes progress against the first but does not require it to have been run through completion, is undertaken by the user community. Specifically, clinical users, biopharma companies, and others less engaged in the design and commercialization of the assays for use, follow the following steps:

1. Statement of Value to stakeholders: patients, manufacturers, Pharma, etc. This should be stated in the context of the alternatives.
2. Implement and refine protocols for the intended use and develop / merge databases from various sources to support validation and qualification of the biomarker.
3. Clinical Performance Groundwork to characterize sensitivity and specificity for readers using the assay. Should include intra-reader test-retest (TRT) across single vendor and single site operating conditions.
4. Clinical Efficacy Groundwork to qualify biomarker as a surrogate endpoint in "real world" imaging conditions, and extending to intra-reader TRT, multi-vendor, multi-site settings.

QIBA is set up to cover both perspectives and offers a process framework and a project management structure that address the incremental progress of various biomarkers at different stages of these activities.

Figures courtesy Aerts, Wahl, and Mozley